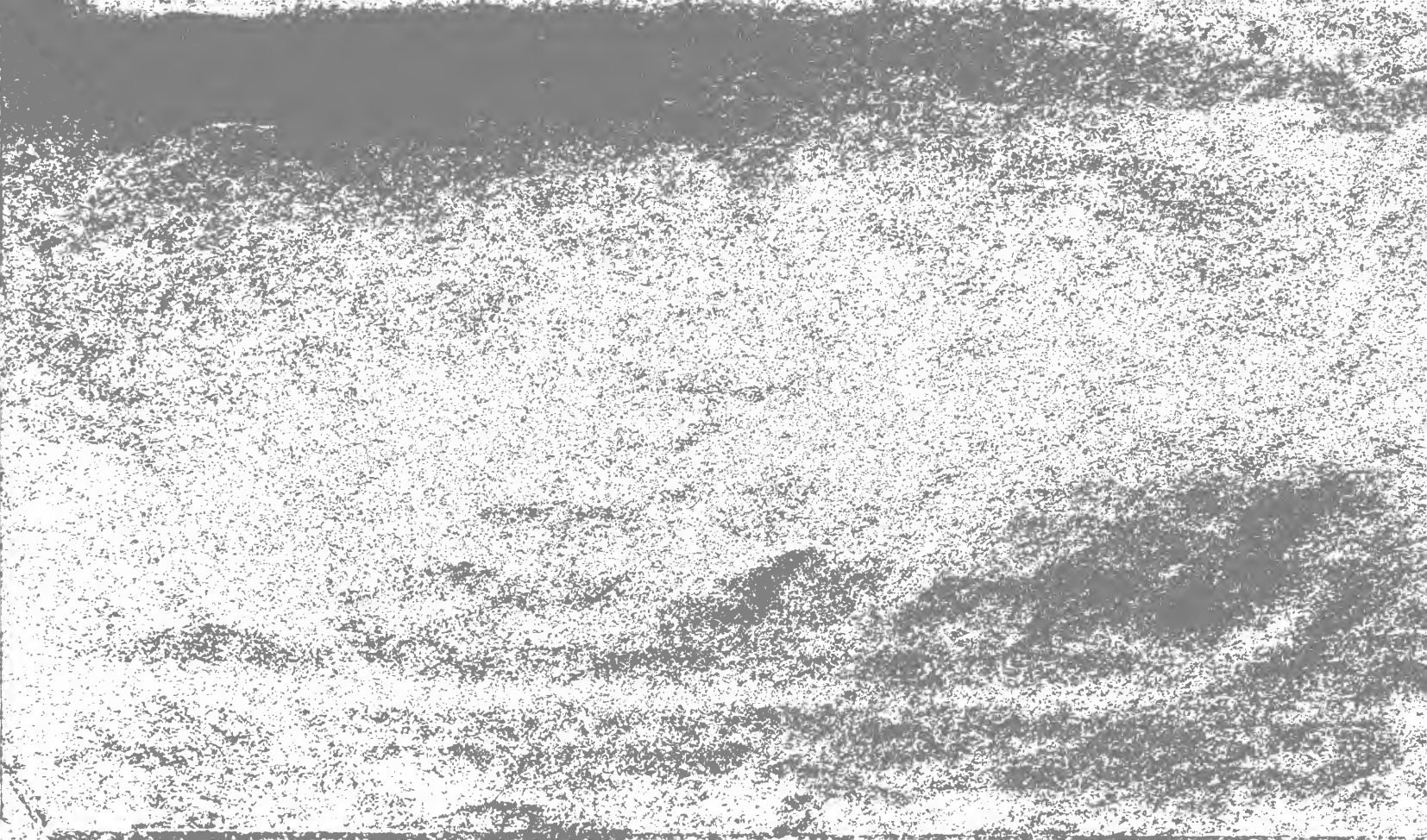

BULLETIN NO. 32

BUREAU OF EDUCATIONAL RESEARCH
COLLEGE OF EDUCATION

THE INTERPRETATION OF THE PROBABLE ERROR AND THE COEFFICIENT OF CORRELATION

By

CHARLES W. ODELL

Assistant Director, Bureau of Educational Research

PRICE 50 CENTS

PUBLISHED BY THE UNIVERSITY OF ILLINOIS, URBANA

1926

TABLE OF CONTENTS

PREFACE

Graduate students and other persons contemplating educational research frequently ask concerning the need for training in statistical procedures. They usually have in mind training in the technique of making tabulations and calculations. This, as Doctor Odell points out, is only one phase, and probably not the most important phase, of needed training in statistical methods. The interpretation of the results of calculation has not received sufficient attention by the authors of texts in this field. The following discussion of two derived measures, the probable error and the coefficient of correlation, is offered as a contribution to the technique of educational research. It deals with the problems of the reader of reports of research, as well as those of original investigators. The tabulating of objective data and the making of calculations from the tabulations may be and frequently is a tedious task, but it is primarily one of routine. The interpreting of the results of calculation is not a routine task. Many conditions affect their meaning and the research worker constantly encounters new problems of interpretation. It is, however, possible to state certain general principles that will serve as a guide in this phase of educational research.

WALTER S. MONROE, *Director.*

July 7, 1926

THE INTERPRETATION OF THE PROBABLE ERROR AND THE COEFFICIENT OF CORRELATION

CHAPTER I

INTRODUCTION

Purpose of this bulletin. One of the most noticeable recent developments in the field of education has been the extensive application of statistical methods to the description of educational conditions and the solution of educational problems. Only a comparatively few years ago conditions were portrayed chiefly in terms of adjectives and other expressions of quality or degree, but now these have been superseded to a considerable extent by definite quantitative terms. There are at least two reasons why everyone engaged in educational work, from the classroom teacher to the research expert, should become acquainted with certain commonly used formulae, methods of computation, and other statistical procedures. In the first place, situations frequently are encountered in which it is desirable to make use of statistical procedures for the purpose of collecting and analyzing data which have a bearing upon practical educational problems. For the great majority of educational workers, however, it is probably more important to be able to interpret correctly the numerous statistical expressions and discussions which are encountered in professional reading and other work. It is almost impossible to peruse a single issue of an educational periodical or a recent book in the field of education or to attend an educational meeting without seeing or hearing many statistical terms employed. Most of the commonly used methods of computation can be mastered by practically any person of average intelligence and arithmetical ability within a rather short time, but the power of interpreting correctly the various measures derived by statistical methods is not so easily acquired. The acquisition of this power demands considerable familiarity with the concepts involved and this in turn requires clear and critical thinking.

It is with the second of the two purposes mentioned in the preceding paragraph chiefly in mind that the writer has attempted in this bulletin to throw some light on the use and interpretation of two of the most frequently used statistical measures,[1] the probable error[2] (com-

[1]The term "statistical measure," sometimes shortened to "measure," is used in this bulletin to refer to a measure or quantitative expression which has been derived from a number of data such as scores or other measurements and which summarizes

monly abbreviated *P.E.*), and the coefficient of correlation (commonly abbreviated *r*), in the hope that readers will be helped in their understanding of the significance of these terms. Since the methods of computing them can be found in many places,[3] their actual calculation will not be explained in detail, although the formulae for them will be given.

---

or expresses in a single numerical index some tendency of the original data. All such expressions as means, medians, modes, measures of deviation, measures of relationship, and so on are statistical measures. In order to avoid confusion the term "measure," which is often used to refer to the result obtained by applying a measuring instrument to an individual case, will not be used in this bulletin in that sense, but "score" or "measurement" will be used instead.

[2]The term "probable error" *(P.E.)* has come to be generally used to include both the probable error proper and the median deviation (abbreviated *Md.D.*), although, as will be shown later in the discussion, the latter is in no real sense an error. For this reason and also to avoid confusion the term probable error will sometimes be used when median deviation would be preferable from the standpoint of strict accuracy of use. The reader should not obtain the idea, however, that the writer believes it is desirable to use probable error instead of median deviation; he distinctly does not believe so.

[3]See:

Odell, C. W. Educational Statistics. New York: The Century Company, 1925, p. 138-39, 150-8-, 221-41, or any other standard text on statistics.

CHAPTER II

THE PROBABLE ERROR[1]

Formula for the probable error. Since the probable error, as shown by the substitute term median deviation, is the median of the deviations or differences of the individual scores or measurements from their average,[2] it may be computed simply by determining the median of these deviations or differences. However, the customary method is to determine the standard deviation[3] first and then to multiply it by .6745[4] to obtain the probable error. In other words the usual formula for the probable error is:

$$P.E. = .6745\sigma.$$

The relationship existing between the probable error and the standard deviation is therefore of the same sort as that existing between a foot and a yard or a pint and a quart, that one always equals the other multiplied by a constant factor. Thus just as .5 quart equals a pint and 2 pints a quart, so $.6745\sigma$ equals 1 *P.E.* and 1.4826[5] *P.E.* equals 1σ.

Different uses of the probable error. There are several more or less different uses or meanings of the probable error, at least five of

---

[1]The probable error, often more properly called the median deviation, is only one of several commonly used measures of the same sort. Among the other similar ones are the standard deviation (abbreviated *S.D.* or σ (sigma)), which in certains uses becomes the standard error, the mean deviation (*M.D.* or *A.D.*), the quartile deviation or semi-interquartile range (*Q*), and the 10-90 percentile range *(D)*. All of these except the last are rather frequently encountered. In general, whatever is said about the probable error may be applied to these other measures also. The one important exception to this statement is that since these measures, except *Q*, differ from the probable error in magnitude, their interpretations in numerical statements will, of course, differ. For a discussion of these other measures see:

. Odell, op. cit., p. 120-38.

[2]The term "average" is used here in a general sense, that is, it includes the arithmetic mean, commonly called the average, the median, the mode, and all other measures of central tendency. Deviations or differences are usually computed from the arithmetic mean but may be taken from any other measure of central tendency.

[3]The formula for the standard deviation is $\sigma = \sqrt{\frac{\Sigma x^2}{N}}$ in which x denotes the deviation or difference of a particular score from the average, N stands for the total number of cases or scores and Σ (sigma) is the symbol for summation.

[4]It is only in the case of a normal distribution or by chance that the probable error is equal to nearly .6745 times the standard deviation. However, most educational data form distributions which approximate normality closely enough that no serious error is involved in using the given decimal as the multiplier.

[5]This number is of course the reciprocal of .6745.

which are fairly distinct from one another, and will be dealt with in this discussion.

1. A measure of the spread or variability of a distribution of data about the average. When used in this way it should properly be called the median deviation (*Md.D.*). If the term probable error is employed it should be followed by the words "of the distribution" and abbreviated $P.E._{Dis.}$.

2. A unit of measurement. This also is a use for which the term median deviation is really the correct one to employ, since it involves merely a particular use of the median deviation of a distribution. Since no subscript has been agreed upon to denote this use, the writer suggests "*U*" for "unit." Thus, when designating the median deviation used as a unit of measurement one should write $Md.D._{U.}$ or if one follows the general practice rather than the best, $P.E._{U.}$.

3. A measure of the reliability of sampling. The accepted abbreviation for this use is *P.E.* with a subscript denoting the measure to which it applies. Thus $P.E._{M.}$ denotes the probable error of the mean, $P.E._{Md.}$ that of the median, $P.E._{r}$ that of the coefficient of correlation and so on.

4. A measure of the reliability or accuracy of any one of a number of scores or measurements of the same thing. As will be explained later this is from one standpoint a variety of the immediately preceding use. It has no conventional abbreviation, hence $P.E._{Sc.}$ is suggested as a suitable one.

5. A measure of the reliability of a measuring instrument. This may be divided into two sub-heads as follows:

A. A measure of the reliability or accuracy of scores obtained from a measuring instrument when compared with those obtained from another application of the same or of a supposedly equivalent measuring instrument to the same individuals. This is called the probable error of estimate and is abbreviated $P.E._{Est.}$.

B. A measure of the reliability or accuracy of scores obtained from a measuring instrument when compared with the theoretically true scores. This is called the probable error of measurement and is best abbreviated $P.E._{Meas.}$.

The probable error as a measure of the spread or variability of a distribution of data around its average. As was stated above the term

probable error is a misnomer in connection with this use and median deviation should be used instead. Therefore, the writer will use the latter expression in the discussion immediately following. The use of the median deviation as a measure of the spread or variability of a distribution of data around its average is the fundamental one and from it all the others are derived. When a number of scores or measurements yielded by a test or other measuring instrument are tabulated in a distribution it is frequently desirable and useful to describe in some concise way their spread or variability about the average. In other words, one often desires to indicate or summarize by a single numerical expression the extent to which the individual scores tend to cluster about or depart from their average. For example, if the marks assigned the pupils in two classes have been tabulated and the averages of both classes are computed and found to be 85 percent, one knows that the average rating of the classes is the same but he does not know whether or not the classes are equally homogeneous in regard to the ratings given. In other words, he does not know whether all the pupils in both classes received marks closely grouped around the average, whether their marks ranged from decidedly below to considerably above the average, or whether the first condition held in one class and the second in the other.

One of the measures most commonly used as an index of the amount of spread or variability is the median deviation.[6] This is exactly what its name implies, the median of the deviations or differences of the individual scores from their average. Since the median is a point on each side of which there are half of the measures in the whole distribution, the median deviation is always of such a magnitude that half of the individual scores differ from their average by less than this amount and half by more. For example, if one of the classes referred to above had a median deviation of 3 percent it would mean that half of the pupils' marks were within 3 percent of 85, that is, from 82 to 88, and the other half either below 82 or above 88. Similarly a median deviation of 5 percent for the other class would mean that the marks of half of its members were between 80 and 90 and those of the other half either below 80 or above 90. From these values of the median deviation, 3 and 5, one would know that the first class was more homogeneous than the second in respect to the ratings given.

---

[6]It cannot be said in any real sense that the differences between the individual scores or measures of a number of individuals and their average are errors. Despite this fact, however, the term probable error is frequently used in this connection.

TABLE I.[6] A SUMMARY OF TABLE I OF JOHNSON'S STUDY GIVING THE MEAN ACCURACY SCORES EARNED ON THE COURTIS SUBTRACTION CARD NO. 33 BY THE GROUPS USING THE SEVERAL METHODS OF SUBTRACTION

Score	Method				
	I	II	III	IV	Mixed
17	75	13	2	8	3
16	74	5	4	3	6
15	35	2	1	1[a]	1
14	22	3		1	2
13	5		1		
12	6				
11	1				1
10	1				
9	1				
N	220	23	8	13	13
M	15.7	16.2	15.8	16.4	15.5
Md.D.	0.9	0.7	0.8	0.6	1.1

[a]Printed as 5 but here taken as 15, since the use of the latter value checks with the mean reported.

The actual use of the median deviation in this way is shown by the following table taken from a magazine article.[7] This table shows the distributions of scores on the Courtis Subtration Card No. 33 made by five groups of pupils who had used different methods of subtraction. Below each column in the table are given the number of pupils, the mean score, the standard deviation and the median deviation of the

[7]Rugh, G. M., Knight, F. B., and Lutes, O. S. "On the relative merits of subtraction methods: another view," Journal of Educational Research, 11:154-55, February, 1925. For other examples of the use of the probable error or median deviation see the following references:

Courtis, S. A. The Gary Public Schools: Measurement of Classroom Products. New York: General Education Board, 1919, p. 213.

Stoddard, G. D. "Iowa Placement Examinations." University of Iowa Studies in Education. Vol. 3, No. 2. Iowa City: University of Iowa, 1925, p. 62-64.

Kallom, A. W. "Times of writing each of the Arabic numerals determined by the reaction time method." Journal of Educational Psychology, 7:226-28, April, 1916.

Childs, H. G. "Measurement of the drawing ability of two thousand one hundred and seventy-seven children in Indiana city school systems by a supplemented Thorndike Scale." Journal of Educational Psychology, 6:391-408, September, 1915.

[6]For purposes of convenience the tables in this bulletin are numbered consecutively instead of as in the sources from which they are quoted. Also some of them have been modified slightly in order to be consistent or to follow the best form, parts of some have been omitted, and occasional errors have been corrected.

distribution in that column. For example, 220 pupils used the first method, their mean score was 15.7, and the median deviation of their scores .9. This statement is merely a way of expressing the fact that half of the scores probably fell within .9 of the mean, or between 14.8 and 16.6, and half outside of these limits. Similarly, for the pupils who used the second method the mean was 16.2 and the median deviation .7, which indicates that half of the pupils probably made scores between 15.5 and 16.9 and half lower or higher than these limits.

It will perhaps be helpful to illustrate the significance of the median deviation by a graphical representation. With this in mind Figure 1 has been prepared. The portion of the figure at the left represents graph-

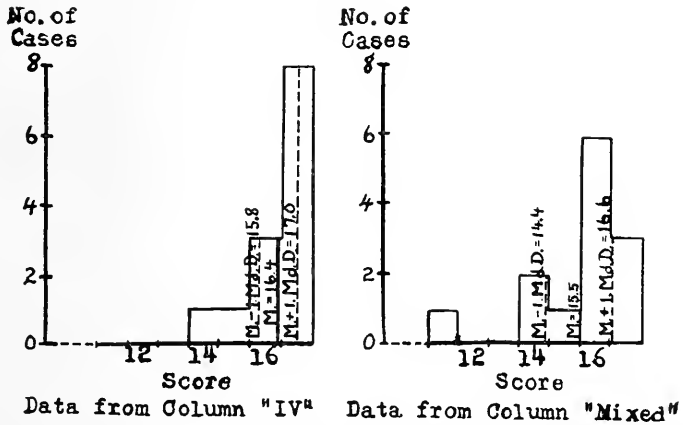

FIGURE 1. GRAPHICAL REPRESENTATION OF THE DATA IN THE LAST TWO COLUMNS OF TABLE I

ically the distribution of scores contained in Column IV of Table I, the portion at the right the scores in the column headed "Mixed." The distributions in these two columns were chosen for graphical representation because the total number of scores in each is the same and therefore the areas of the surfaces representing them are equal. Inspection of the figure makes it evident that the scores represented at the right spread out considerably more than do the others. The height of the graph at the right is less and the length of its base greater than of the one at the left, which indicates a wider spread of scores. This agrees with the fact that the median deviation of the distribution represented by it is 1.1, whereas that of the other one is only .6. It might be noted also that neither of the graphs approach normality very closely, the one at the right, however, doing so more nearly than the one at the left.

The interpretation of the median deviation, when used to measure how closely individual scores or measurements cluster about their average or how far they spread out from it, may be extended further than has been suggested in the preceding paragraphs by stating what fraction of the scores will not differ from the average by more than a given multiple of the median deviation. For the few smallest integral multiples we may state as follows:[9]

50.00 percent of scores differ from the average by less than 1 *Md.D.*
82.26 percent of scores differ from the average by less than 2 *Md.D.*
95.70 percent of scores differ from the average by less than 3 *Md.D.*
99.30 percent of scores differ from the average by less than 4 *Md.D.*
99.92 percent of scores differ from the average by less than 5 *Md.D.*

We may also change the form of statement and say that the chances are:

1 to 1 that a score differs from the average by less than 1 *Md.D.*
4.6 to 1 that a score differs from the average by less than 2 *Md.D.*
22 to 1 that a score differs from the average by less than 3 *Md.D.*
142 to 1 that a score differs from the average by less than 4 *Md.D.*
1,340 to 1 that a score differs from the average by less than 5 *Md.D.*[10]

---

[9]Although the numerical interpretations given in the text hold exactly only in the case of normal frequency distributions they may be used without serious error in dealing with the large majority of tabulations of such educational facts as pupils' heights, weights, school marks and test scores, teachers' salaries, numbers of pupils to the room, and so forth. For example, 109 of the scores in the first column of Table I fall within 1 *Md.D.* of the mean, whereas 110 would be expected to do so.

[10]It has been previously stated that the chief difference between the interpretation of the standard deviation (σ), the mean deviation (*M.D.*), and the 10-90 percentile range (*D.*), and of the median deviation has to do with numerical interpretation. For example, it is to be expected that:

68.27 percent of scores differ from the average by less than 1 σ
95.44 percent of scores differ from the average by less than 2 σ
99.74 percent of scores differ from the average by less than 3 σ
99.99 percent of scores differ from the average by less than 4 σ

Using the other form of statement, the chances are:

2.15 to 1 that a score differs from the average by less than 1 σ
21 to 1 that a score differs from the average by less than 2 σ
369 to 1 that a score differs from the average by less than 3 σ
15,772 to 1 that a score differs from the average by less than 4 σ

Also it is probable that:

57.51 percent of scores differ from the average by less than 1 *M.D.*
88.94 percent of scores differ from the average by less than 2 *M.D.*
98.33 percent of scores differ from the average by less than 3 *M.D.*
99.86 percent of scores differ from the average by less than 4 *M.D.*

Or the chances are:

1.35 to 1 that a score differs from the average by less than 1 *M.D.*
8 to 1 that a score differs from the average by less than 2 *M.D.*

These more extended interpretations may be illustrated by referring back to the examples used earlier. For the first of the two classes referred to, which had a mean score of 85 and a median deviation of 3, it is not only probable that half of its members have scores between 82 and 88 but also that about 82 percent of them have scores between 79 and 91 (85 ± 6), almost 96 percent between 76 and 94 (85 ± 9), over 99 percent between 73 and 97 (85 ± 12), and very nearly 100 percent between 70 and 100 (85 ± 15). Using the other form of statement for the first column of Table I, the chances are 1 to 1, or even, that a particular score chosen at random falls between 14.8 and 16.6 ($15.7 \pm .9$), 4.6 to 1 that it falls between 13.9 and 17.5 (15.7 ± 1.8), 22 to 1 that it is between 13.0 and 18.4 (15.7 ± 2.7), 142 to 1 that it is between 12.1 and 19.3 (15.7 ± 3.6), and 1340 to 1 that it is between 11.2 and 20.2 (15.7 ± 4.5).[11]

The probable error as a unit of measurement.[12] In dealing with data of various sorts one encounters many different units. The unit usually used for school marks is the percent, for ages the year, the month, or the day, for salaries the dollar, for heights the foot or the inch, for weights the pound, for spelling the word, for arithmetic the example, and so on. In the case of such characteristics as height, weight, age, salary, and so forth, even though there are commonly used units of measurement, it is difficult if not impossible to compare one trait with another. For example, one cannot readily determine whether a pupil's height of four feet, eleven inches, his weight of 102 pounds, or his age of 12 years and 8 months is the highest or lowest ranking when

---

59 to 1 that a score differs from the average by less than 3 *M.D.*
706 to 1 that a score differs from the average by less than 4 *M.D.*

For the 10-90 percentile range the corresponding statements are:

99 percent of scores differ from their average by less than 1 *D.*
99.99997 percent of scores differ from their average by less than 2 *D.*

And the chances are:

95 to 1 that a score differs from the average by less than 1 *D.*
3,380,614 to 1 that a score differs from the average by less than 2 *D.*

As was suggested previously the quartile deviation may be interpreted in the same way numerically as the median deviation.

[11]The fact that one, or sometimes even both, of the limits within which a certain fraction of the scores may be expected to fall comes outside the range of actually obtained scores is due to the fact that the scores do not form a normal distribution. That they do not is often caused by the number of scores being small, as well as by causes inherent in the nature of the data themselves.

[12]This use is derived directly from the one discussed in the preceding paragraph and also is one to which the name median deviation should properly be applied. The writer will therefore employ the latter term, abbreviated *Md.D.*, throughout his treatment of this use.

compared with other similar pupils. There are also many situations in which there is no commonly used unit or indeed any conventional unit closely connected with the type of thing being measured. Probably most of such cases in the field of education have to do with the measurement of difficulty, such as difficulty of examples in arithmetic, of questions in history or geography, of passages in reading, of words in spelling, and so forth.

To meet the need for a common unit in which all scores and measurements, including those for which no conventional units are available, may be expressed and thereby easily compared the median deviation has been adopted and come into rather common use. Irrespective of the units in terms of which scores or measurements have been expressed originally, by applying certain statistical procedures they may be expressed in terms of median deviations.

The most frequent use of the median deviation as a unit has probably been in connection with the construction of standardized educational measuring instruments. The values or difficulties assigned the different items or steps on the scale or the distances between the steps are very frequently expressed in such units. An example of this may be found in connection with Woody's Arithmetic Scales,[13] given as part of his account of the derivation of these scales. Woody describes how the difficulty values of the exercises composing each scale were determined. The essential steps in this determination consisted of finding the median deviation of the distribution of scores[14] for each scale and then measuring the distance of each exercise from the average of the distribution in terms of *Md.D.* units.[15] Different results were obtained in the different school grades so that it was necessary to combine these into average results. Finally, Woody located zero[16] points, that is, points of absolute

---

[13]Woody, Clifford. "Measurements of Some Achievements in Arithmetic." Teachers College Contributions to Education, No. 80. New York: Teachers College, Columbia University, 1916, p. 29-54.

[14]It is assumed that the distribution of pupils' scores represents the distribution of their abilities.

[15]It does not seem necessary for the purpose of the present discussion to explain in complete detail just how this was done. Briefly, Woody found the percent of pupils obtaining the correct answer to each exercise and, on the assumption that the distribution of pupils' abilities was normal, calculated the degree of difficulty of an exercise in terms of the number of *Md.D.* units that the ability required to do each exercise differed from the average ability of the group. For a fuller explanation of the method of procedure, see:

Odell, op. cit., p. 313-15.

[16]The determination of such zero points is not a necessary part of the process of employing $Md.D._{U.}$ but merely renders the values so expressed more usable. The actual

lack of ability to solve exercises in the four fundamental operations and transformed the $Md.D._{U.}$ values of the various exercises from distances from the averages of the distributions into distances from the zero points. To illustrate this simply, we may express John's height by saying that he is six inches taller than Paul. If, however, we know that Paul is five feet and three inches tall we can express John's height much more satisfactorily for most purposes by saying that he is five feet and nine inches above the zero point which is, of course, zero inches or no height at all.

To show the final result of the process, that is, the difficulty values determined for the exercises, a portion of one of Woody's tables[17] is given as Table II. Exercise 2 was found to be the easiest, having a difficulty value of 1.23 $Md.D._{U.}$, exercise 3 was next with a value of 1.40 $Md.D._{U.}$ and so on up to exercise 38, the most difficult, which had a value of 9.19 $Md.D._{U.}$. After the difficulty values have been so expressed we can not only say, for example, that exercise 3 is .17 $Md.D._{U.}$ and exercise 5 1.27$Md.D._{U.}$ more difficult than No. 2, but also, if the zero point has been located accurately, that exercise 5 is about twice as hard as No. 2, but only half as difficult as No. 15.

The preceding discussion has used the term $Md.D._{U.}$ but perhaps not made clear just what it really means. Since 50 percent of the scores in a normal distribution fall within 1 *Md.D.* of the average and since a normal distribution is symmetrical it follows that half of these 50 percent, or 25 percent, of the scores will fall within 1 *Md.D.* of the average on each side. That is, 25 percent will fall between the average and 1 *Md.D.*

---

determination of zero points usually rests, at least in part, upon opinion as to just what constitutes absolute lack of ability in a given field. Sometimes it is possible to determine rather accurately just what is the least difficult task of a certain sort and to locate that degree of ability just barely insufficient to accomplish this task, but in many cases this can not or at least has not been done.

[17] Woody, op. cit., p. 54. Other examples of the use of $Md.D._{U.}$ as a unit may be found in:

Buckingham, B. R. "Spelling Ability—Its Measurement and Distribution." Teachers College Contributions to Education, No. 59. New York: Teachers College, Columbia University, 1913, p. 40-65.

Monroe, W. S. An Introduction to the Theory of Educational Measurements. Boston: Houghton Mifflin, 1923, p. 61-62, 94-103, 138-41, 150-52.

Trabue, M. R. "Completion-Test Language Scales." Teachers College Contributions to Education, No. 77. New York: Teachers College, Columbia University, 1916, p. 29-73.

Hughes, J. M. "The use of tests in the evaluation of factors which condition the achievement of pupils in high school physics," Journal of Educational Psychology, 16: 217-31, April, 1925.

TABLE II. FINAL VALUES OF ADDITION EXERCISES

No. of Exercise	Value	No. of Exercise	Value	No. of Exercise	Value	No. of Exercise	Value
2	1.23	14	3.92	22	6.44	35	7.97
3	1.40	9	4.18	19	6.79	29	8.04
5	2.50	12	4.19	23	7.11	31	8.18
7	2.61	13	4.85	34	7.43	24	8.22
6	2.83	15	4.97	26	7.47	36	8.58
8	3.21	17	5.52	30	7.61	37	8.67
1	3.26	16	5.59	27	7.62	33	8.67
4	3.35	18	5.73	25	7.67	38	9.19
10	3.63	20	5.75	28	7.71		
11	3.78	21	6.10	32	7.71		

below the average and another 25 percent between the average and 1 *Md.D.* above it. Furthermore the average of a symmetrical distribution falls at the middle of the distribution, so that 50 percent of the scores lie below it and 50 percent above it. Therefore, it is easily seen that 75 percent of the scores lie below 1 *Md.D.* above the average, as this is simply the sum of the 50 percent below the average and the 25 percent between the average and 1 *Md.D.* above it. To make this clearer the accompanying figure is given. The portion of the normal frequency

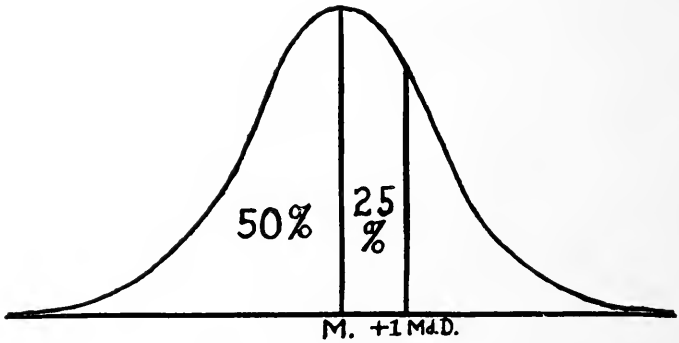

FIGURE 2. REPRESENTATION OF A NORMAL DISTRIBUTION OF SCORES SHOWING THE MEANING OF THE MEDIAN DEVIATION AS A UNIT OF DIFFICULTY

surface to the left of the vertical line at its center, marked *M.*, is the 50 percent of the area below the average. That part between this vertical line and the one erected at + 1 *Md.D.* is the 25 percent between the

average and one median deviation above the average. Thus, all the area to the left of the shorter vertical line is 75 percent of the whole area. With this in mind we can now explain the meaning of the median deviation as a unit of difficulty by saying that it is the difference in difficulty between an exercise answered correctly by 50 percent of the pupils tested and another answered correctly by 75 percent of the pupils.[18] Looking at Table II we see that exercise 25 has a value of 7.67 and exercise 37 of 8.67, a difference of 1.00 $Md.D._{U.}$. We know, therefore, that if the two exercises were given to the same group of pupils and 50 percent of them answered exercise 37 correctly, 75 percent might be expected to answer exercise 25 correctly, since it is 1 $Md.D._{U.}$ easier than the former.

There is also another somewhat different meaning which is often attached to the median deviation when used as a unit of measurement. In the construction of such measuring instruments as handwriting and drawing scales, one method of determining the value or merit of the specimens being rated for a scale is to have them compared with one another by a number of supposedly competent judges. For example, judges compare specimen A with B, also A with C, B with C, and so on. Record is made of how many or what percent of the judges rate A as better than B and of course how many rate B as better than A, and so on. When 75 percent of the judges rate one specimen as better than another[19] the difference in merit between the two is assumed to be 1 $Md.D._{U.}$. This is illustrated by Figure 3 in which the surfaces under the two curves are assumed to represent distributions of judges' ratings of two specimens, A and B. It is assumed that the opinions of judges concerning the merit or value of a specimen will form a normal distribution, the center or average of which is the true value. Therefore, the surface at the left, under curve A, is taken as representing the distribution of judges' opinions concerning specimen A and the point A on the base line where the solid vertical line meets it as the true value of A. Similarly, point B at the foot of the broken vertical line is assumed to represent the true value of B. If 75 percent of the judges rate B as

[18]It is also possible to say that 1 $Md.D._{U.}$ is the difference in difficulty between an exercise answered correctly by 25 percent of the pupils and one answered correctly by 50 percent of them, but the form of statement given above is more usual.

[19]In rating specimens for the purpose being discussed, judges are expected to rate each as better or worse than each of those with which it is compared. If they rate two as equal, the rating must be thrown out or divided between the two. Therefore if 75 percent of judges rate one specimen as better than another, 25 percent must rate it as worse.

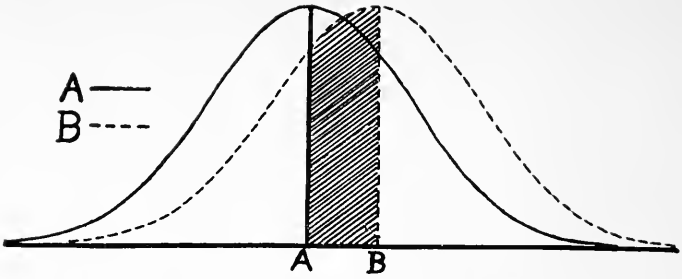

FIGURE 3. ILLUSTRATION OF METHOD OF DETERMINING DIFFERENCE IN MERIT OF TWO SPECIMENS BY JUDGES' RATINGS OF ONE AS BETTER OR WORSE THAN THE OTHER

superior to A, 75 percent of the area of the surface to the right, representing B, will lie above or to the right of the vertical line assumed to show the true value of A and of course 25 percent below or to the left of that line. Since 50 percent of the judges' ratings of B lie above its average merit, that is, to the right of the broken vertical line above point B, the portion of the surface representing B which is included between the two vertical lines must be 75 percent minus 50 percent, or 25 percent. To make clear which this is, it has been shaded in the figure. We have already seen that a distance of 1 median deviation in one direction from the average distribution includes 25 percent of the total number of cases. Therefore, the distance between the two vertical lines must be 1 *Md.D.* in order that 25 percent of the area be included.

This method of determining the value of merit of specimens has been made use of in the case of a number of our standardized scales. Probably the best known example of its use is in connection with Thorndike's Handwriting Scale.[20] In his account of its construction he describes two methods, one of which is that just mentioned. He had samples of handwriting rated by a number of judges as to whether they were better

---

[20] THORNDIKE, E. L. "Handwriting," Teachers College Record, 11:1-41, March, 1910. For further examples, see:

HOKE, E. R. "The Measurement of Achievement in Shorthand." The Johns Hopkins University Studies in Education, No. 6. Baltimore: Johns Hopkins Press, 1922, p. 33-34.

HILLEGAS, M. B. "Scale for the measurement of quality in English composition by young people," Teachers College Record, 13:1-54, September, 1912.

MURDOCH, KATHERINE. "The Measurement of Certain Elements of Hand Sewing." Teachers College Contributions to Education, No. 103. New York: Teachers College, Columbia University, 1919, p. 22-26.

or worse than the other samples and, according to the method outlined above, determined the differences in merit between the samples in terms of the median deviation. A sample considered to possess no merit as handwriting, though obviously an attempt to write, was used as the zero point and the distance of each sample above this point determined.

Probable errors of sampling. The third use of the probable error is one to which that term is properly applied. In this case it is employed directly as a measure of the size of the errors involved in sampling, that is to say, as a measure of the reliability of sampling. The probable error of sampling can not be used alone, but must always be connected with some other measure such as an average, a standard or quartile deviation, a difference, a coefficient or ratio of correlation, a regression coefficient, or other similar measures. Assuming that the sample has been selected in a random manner, in other words that it is not biased, the probable error of sampling gives an indication of how reliable such derived measures are when the cases upon which they are based are considered as a sample of a larger number of similar ones. For example, if the average score of five hundred eighth-grade children upon an intelligence test has been determined and it is assumed that no errors are present in the test scores or computations leading to the average, this average is the true one for the children actually tested. If the five hundred children have been selected from a much larger number in a city school system, the average obtained from their scores is not, except by chance, the true average of all the eighth-grade children in the system. However, if we assume that the five hundred children constitute a random sample, we can determine the reliability of the average actually obtained when considered as the average of all of the eighth grade children in the system.

When the probable error of sampling is used, it is both customary and convenient to place a plus and minus sign, followed by the probable error, immediately after the measure to which it applies. Thus if the average intelligence quotient of the five hundred pupils had been found to be 102 and its probable error 3, it would frequently be written 102 ± 3, when considered as an average *I.Q.* of all of the eighth-grade pupils in the system. The same practice is also followed in the case of other measures than the average. A second fairly common way of referring to the probable error of sampling is to use the abbreviation *P.E.* with a subscript indicating the measure to which it applies. Thus $P.E._{M.}$ denotes the probable error of the mean, $P.E._{Md.}$ that of the median, $P.E._{r}$ that of the coefficient of correlation, and so on.

The interpretation of the probable error of sampling from the standpoint of chance is the same as that of the median deviation when used as a measure of variability or scatter. That is, the chances are even that the true measure of the whole group does not differ from the measure obtained from the sample by more than the value of the probable error; they are 4.6 to 1 that it does not differ by more than 2 *P.E.*, 22 to 1 that it does not differ by more than 3 *P.E.*, and so on. Another way of stating the same thing is that if a number of samples of the same size as the one already taken and similar to it were selected and corresponding measures computed from them, half of these measures would probably fall within 1 *P.E.* of the first one computed, 82 percent within 2 *P.E.*, 96 percent within 3 *P.E.*, and so on. Thus, in the case of the group of eighth-grade pupils referred to, it is probable that, if a number of similar samples were chosen and their means determined, half of them would fall within 3 points of 102, that is between 99 and 105, 82 percent between 96 and 108 (102 ± 6), 96 percent between 93 and 111 (102 ± 9), and so forth.

A good example of this use of the probable error is to be found in a recent issue of the Journal of Educational Psychology.[21] In the article referred to the following table is given. It contains a number of means, standard deviations, and coefficients of correlation, each followed by its probable error. For example, the mean English grade of the first high-school group is given as $84.6 \pm .3$. This indicates that if similar samples were taken it is probable that half of the obtained means would be between 84.3 and 84.9 ($84.6 \pm .3$), 82 percent of them between 84.0 and 85.2 ($84.6 \pm .6$), 96 percent between 83.7 and 85.5 ($84.6 \pm .9$), and so on.

The formulae by which to compute a probable error of sampling differ according to the measure for which it is being found. The follow-

---

[21]Gowen, John W., and Gooch, Marjorie. "The mental attainments of college students in relation to previous training," Journal of Educational Psychology, 16:547-68. November, 1925. Other examples may be found in the following references:

Rich, S. G., and Skinner, C. E. "Intelligence among normal school students." Educational Administration and Supervision. 11:639-44, December, 1925.

Ellis, R. S. "A comparison of the scores of college freshmen and seniors on psychological tests," School and Society, 23:310-12, March 6, 1926.

Remmers, H. H., and Edna M. "The negative suggestion effect of true-false examination questions," Journal of Educational Psychology, 17:52-56. January. 1926.

Monroe, W. S. An Introduction to the Theory of Educational Measurements. Boston: Houghton Mifflin Company, 1923, p. 204.

TABLE III. RELATION OF GRADES IN COLLEGE SUBJECTS TO HIGH SCHOOL ENGLISH

High school grades				College grades			
Subject	Average	*S. D.*	*r*	Subject	Average	*S. D.*	*N*
English	84.6 ± .3	7.95 ± .16	.284 ± .026	English	81.5 ± .3	9.12 ± .18	563
English	84.2 ± .2	7.86 ± .17	.209 ± .030	Chemistry	80.5 ± .3	9.18 ± .20	471
English	83.8 ± .3	8.31 ± .21	.220 ± .033	Algebra	82.1 ± .4	11.59 ± .29	361
English	83.7 ± .3	8.33 ± .22	.276 ± .033	Anal. Geom	82.2 ± .4	10.34 ± .27	342
English	85.3 ± .8	7.16 ± .56	.403 ± .092	Elem. French	81.6 ± 1.1	10.16 ± .80	37
English	86.5 ± .5	7.88 ± .33	.389 ± .050	Adv. French	86.9 ± .5	9.13 ± .38	129
English	84.4 ± .4	8.01 ± .25	.190 ± .041	German	82.7 ± .5	9.59 ± .30	240

ing are the formulae for the probable errors of the mean, the median, the standard deviation, and the coefficient of correlation:

$$P.\,E._{M} = \frac{Md.\,D.}{\sqrt{N}}$$

$$P.\,E._{Md.} = 1.2533\,\frac{Md.\,D.}{\sqrt{N}}$$

$$P.\,E._{\sigma} = .7071\,\frac{Md.\,D.}{\sqrt{N}}$$

$$P.\,E._{r} = \frac{1 - r^2}{\sqrt{N}}$$

In these typical formulae it will be noticed that there is one common element, $\sqrt{N}$, appearing in their denominators. The same is true of the formulae for the probable errors of sampling of almost all commonly used statistical measures. Even in those cases in which $\sqrt{N}$ does not appear directly in the denominator of the formulae, it or some similar expression is usually in some way contained therein. Since N stands for the number of cases in the sample, it can easily be seen that the larger the sample the larger is the denominator of the fraction and therefore the smaller the value of the probable error. In other words, increasing the size of the sample decreases the size of the probable errors present and hence increases the reliability or accuracy of the derived measures.

The probable error of a number of measurements of the same thing. Another situation in which the term probable error is appropriate is in measuring the size of the variable errors present in a number of measurements of the same thing. In most, if not all, situations it is impossible to measure a trait with such a high degree of precision and reliability that all similar measurements will agree exactly with the original one. For example, let us suppose that ten different persons determine a child's height or that the same person does so ten times. If height is being found only to the nearest inch and the persons doing the measuring are fairly competent it is likely that all the results will agree. If, however, the attempt is made to secure a rather high degree of accuracy and results are given, let us say, to the nearest sixteenth of an inch, it is extremely improbable that the results obtained, whether by ten different persons or by the same person at ten different times, will be the same. There are generally two causes for this and frequently a third one. In the first place, even though the persons making

the measurements are reasonably competent it is unlikely that all have just the qualities, such as keenness of eyesight, steadiness of hand, ability to time accurately, and so forth, necessary for accurate measurement or that all exercise exactly the same degree of care. Secondly, it is improbable that the child being measured will assume exactly the same posture when all ten measurements are being made. In addition, if different measuring instruments are used it is very unlikely that they are absolutely identical. The errors due to all these and any other chance causes are called variable errors[22] and are often measured by the probable error. In a sense they may be thought of as errors of sampling, for just as a group too large to have all its members measured is sampled by measuring a part of them, so a characteristic which cannot be measured with absolute accuracy and therefore theoretically requires that an infinite number of measurements be made and averaged to secure a perfect one, is sampled by making a limited number of measurements.

A common example of the occurrence of variable errors is in connection with the giving of written examinations and tests. At one testing period a pupil may happen to be feeling unusually well, whereas at another his health may be below par; at one time he may have reviewed the material covered by the questions recently, but at another it may happen that the questions touch material about which he knows little although he remembers most of what he has studied; at one time he may make a better score than he deserves by cheating, whereas at another his score may not indicate his true ability because his pencil broke or something outside the window attracted his attention, and so on. Similarly, when weight is being measured the result will vary according to whether or not the individual has eaten a meal recently, whether he is wearing heavier or lighter clothing than usual, has more or less in his pockets, and so on.

Since, because of all these variable errors, we can rarely, if ever, establish that any one obtained score is a true or even the best obtainable measurement of the characteristic being dealt with, the best that we can do is to supplement the scores obtained by a statement of their reliability. As in the case of the probable error of sampling so here the *P.E.* is commonly affixed to the obtained measure with a plus or minus sign connecting the two. Thus, a pupil's height may be stated as

---

[22]For a fuller discussion of variable errors see:

Monroe, W. S. "The constant and variable errors of educational measurements." University of Illinois Bulletin, Vol. 21, No. 10, Bureau of Educational Research Bulletin No. 15. Urbana: University of Illinois, 1923. 30 p.

TABLE IV. DISTRIBUTIONS OF TEACHERS' MARKS OF PAPERS "A," "B" AND "C"

Mark	A Pass. 75[a]	A Pass. 70	A U. of W. Pass. 70	A U. of C Pass. 70	B Pass. 75	B Pass. 70	B U. of W. Pass. 70	B U. of C. Pass. 70	C Pass. 75
25–29									1
30–34									
35–39									1
40–44									2
45–49									2
50–54									6
55–59					1	1			8
60–64	2			1	1	2		1	17
65–69	1	1			6	6	2	10	19
70–74	2	1	1	5	5	11	7	14	13
75–79	5	6	1	4	24	9	7	20	27
80–84	18	7	6	24	27	13	27	21	11
85–89	24	17	16	31	19	5	24	23	7
90–94	30	15	40	25	7	3	18	9	2
95–100	9	4	22	7	1	1	1		
Total	91	51	86	97	91	51	86	98	116
Median	88.1	87.6	91.9	86.8	81.1	77.6	84.5	80.5	70.3
P.E.	4.9	4.4	3.4	4.4	4.7	5.7	4.3	5.1	8.0

[a]The expression "Pass. 75" indicates that the group of marks in the column under that heading was given by teachers in schools which had a passing mark of 75. Likewise "Pass. 70" indicates that the marks were given by teachers in schools having a passing mark of 70.

62.00 ± .25 inches, which, of course, means that if it is measured a number of times half of the measurements will probably fall between 61.75 and 62.25 inches, 82 percent of them between 61.50 and 62.50 inches, and so on. An example of this use of the probable error (*P.E.*) is shown in Table IV, taken from Kelly,[23] who made use of data collected by Starch and Elliott.[24] This table shows the marks given to each of three papers by several groups of teachers. The first column, for example, shows that of ninety-one teachers in schools with passing marks of 75 two rated paper A from 60 to 64, inclusive, one rated it between 65 and 69, two between 70 and 74, and so on. The median rating was 88.1 and the probable error of the ratings 4.9. In other words, half of them were within 4.9 points of 88.1 and half were not.

To make the situation more concrete Figure 4 has been prepared.

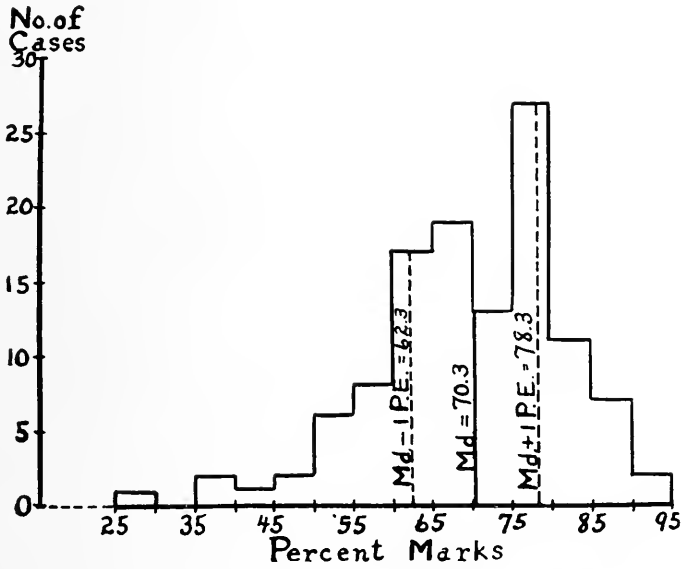

FIGURE 4. GRAPHICAL REPRESENTATION OF MARKS GIVEN IN THE LAST COLUMN OF TABLE IV

[23]KELLY, F. J. "Teachers Marks: Their Variability and Standardization." Teachers College Contributions to Education No. 66. New York: Teachers College, Columbia University, 1914, p. 55. Also see:

TRABUE, M. R. Measuring Results in Education. New York: American Book Company, 1924, p. 199-203, 259-67.

[24]STARCH, D., and ELLIOTT, E. C. "Reliability of the grading of high school work in English," School Review, 20:442-57, September, 1912.

STARCH, D., and ELLIOTT, E. C. "Reliability of grading work in mathematics," School Review, 21:254-59, April, 1913.

It represents graphically the distribution of marks shown in the last column of Table IV. At a distance of 1 *P.E.*, that is 8.0, from the median, which is 70.3, vertical lines have been erected. It can be seen by rough inspection that approximately half of the area of the column diagram lies between these two lines or, in other words, half of the marks fall within these two lines and, of course, the other half outside of them.

The best estimate of the true rating of the paper is 70.3, the median, with a *P.E.* of 8.0., which means that half of the marks given by the group of 116 persons who rated the paper probably fall within 8.0 points of 70.3, that is, between 62.3 and 78.3, 82 percent of them fall between 54.3 and 86.3, and so on. The same may also be expected of marks given by other raters equally competent with those in the first group.

Errors of estimate and measurement. Since none of our standardized tests and scales of other measuring instruments are perfectly reliable, that is, since two or more applications of the same instrument or of supposedly equivalent forms thereof cannot be relied upon to yield exactly the same measurements, there are evidently some errors involved. When the data consist of two series of scores or measurements of a number of individuals[25] rather than of a number of measurements of one individual, the errors involved are known as errors of estimate and of measurement. Since these tend to form normal distributions, as do all other variable errors, the median deviation of their distribution may be used as a measure of their magnitude. When this is done the terms probable error of estimate and probable error of measurement are applied. These are commonly abbreviated $P.E._{Est.}$ and $P.E._{Meas.}$ Occasionally, instead of $P.E._{Est.}$ one finds $P.E._{Score}$ and instead of $P.E._{Meas.}$, $P.E._{M.}$. The writer recommends, however, that these latter abbreviations not be used since they might be confused with those for other uses of the probable error.

---

[25]That is, when each individual has been measured twice. The first measurements of all individuals constitute one series and the second ones the other. Such data as these or any other which are correlated are frequently called "sets of paired facts."

The most commonly used formula[26] for these measures are as follows:

$$P.\,E._{Est.} = .6745\sigma\sqrt{1 - r_{12}^2} \text{ or } Md.\,D.\sqrt{1 - r_{12}^2} \text{ and}$$

$$P.\,E._{Meas.} = .6745\,\frac{\sigma_1 + \sigma_2}{2}\sqrt{1 - r_{12}} \text{ or } \frac{Md.\,D._1 + Md.\,D._2}{2}\sqrt{1 - r_{12}}.$$

In the first, the σ or $Md.D.$ used is that of the distribution of which the scores are being estimated. That is, if scores in the second series are being estimated from those in the first, the σ or $Md.D.$ of the second is used. In the second, the two σ's or $Md.D.$'s are averaged. The r_{12} in both is the coefficient of correlation between the two series of scores.

As was suggested above, the probable error of estimate is a measure of the differences between the results obtained by measuring a group of individuals a first and a second time with the same or similar instruments. Occasionally, this definition is extended to include the differences between any two series of scores of the same individuals if they are used for purposes of predicting or estimating one in terms of the other. The probable error of measurement differs in that it measures the differences between the scores obtained from one of the two series of measurements and the theoretically true scores of the individuals tested. Since the theoretically true scores are the averages of infinite numbers of obtained scores with practice effects and all other constant errors eliminated, it is impossible to determine them. If two series of scores are available, however, it is possible to compute measures of the size of the differences between a series of actually obtained scores and the theoretically true ones. The probable error of estimate is always larger than the probable error of measurement because the differences between two series of scores, both of which contain variable errors, are naturally greater than those between a series of scores containing variable errors and another series of theoretically true scores which contain no such errors.

Table V, taken from a critical study of a standardized test,[27] gives certain results obtained from giving two intelligence scales, the Illinois and the National, to several thousand pupils. The number of pupils

---

[26]Several other formulae are sometimes employed, especially for the probable error of measurement. See:

Odell, op. cit., p. 230-41.

[27]Monroe, Walter S. "The Illinois Examination." University of Illinois Bulletin, Vol. 19, No. 6, Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921, p. 58.

TABLE V. CORRELATION BETWEEN SCORES YIELDED BY ILLINOIS GENERAL INTELLIGENCE SCALE AND BY NATIONAL INTELLIGENCE SCALE

Grade	Number of Cases	r	$P.E._{Est.}$	$\frac{P.E._{Est.}}{\text{Average}}$
III A	357	0.53	9.1	0.22
IV B	416	0.70	9.6	0.18
IV A	335	0.74	8.0	0.14
V B	460	0.55	8.7	0.14
V A	285	0.47	12.0	0.19
VI B	383	0.44	12.6	0.17
VI A	259	0.67	10.8	0.13
VII B	350	0.70	11.0	0.12
VII A	210	0.68	10.3	0.11
VIII B	271	0.72	10.2	0.10
VIII A	289	0.69	10.9	0.10
All Grades	3615	0.81	11.5	0.16

tested, the coefficient of correlation, the probable error of estimate and the ratio of $P.E._{Est.}$ to the average are given for each half-grade from IIIA to VIIIA, inclusive, also for all combined. The meaning of the $P.E._{Est.}$ in the last line, for example, is that if probable scores of pupils in Grades IIIA to VIIIA upon the Illinois General Intelligence Scale are estimated from actual scores upon the National Intelligence Scale, half of them will be in error by less than 11.5 points and half of them by more than that amount. Furthermore, 82 percent will be in error by less than twice 11.5 or 23, 96 percent by less than 34.5, and so on.

An example of the use of the probable error of measurement may also be cited from the same source.[28] Table VI gives, for Grades III

[28] MONROE, op. cit., p. 49. Other illustrations may be found by consulting the following sources:

MONROE, WALTER S. "A Critical Study of Certain Silent Reading Tests." University of Illinois Bulletin, Vol. 19, No. 22, Bureau of Educational Research Bulletin No. 8. Urbana: University of Illinois, 1922, p. 33-34.

MONROE, W. S., DEVOSS, J. C., and KELLY, F. J. Educational Tests and Measurements, Revised and Enlarged. Boston: Houghton Mifflin Company, 1924, p. 410.

THORNDIKE, E. L., and SYMONDS, P. M. "Difficulty, reliability, and grade achievements in a test of English vocabulary," Teachers College Record, 34:438-45, November, 1923.

DEARBORN, W. F. "Reliability and uses of group tests of intelligence." Eleventh Conference on Educational Measurements. Bulletin of the School of Education, Indiana University, Vol. I, No. 3. Bloomington: Indiana University, 1925, p. 115-30.

TABLE VI. PROBABLE ERRORS OF MEASUREMENT AND RATIO OF MEASUREMENT TO AVERAGE SCORES.

Grade	Forms	Intelligence		Arithmetic		Silent Reading			
						Comprehension		Rate	
		$P.\ E._M$[a]	$\frac{P.\ E._M}{Av.}$	$P.\ E._M$	$\frac{P.\ E._M}{Av.}$	$P.\ E._M$	$\frac{P.\ E._M}{Av.}$	$P.\ E._M$	$\frac{P.\ E._M}{Av.}$
III	1 with 2	3.5	0.10	2.6	0.17	1.2	0.16	13.7	0.12
	1 with 3					1.3	0.17	13.7	0.11
	2 with 3					1.3	0.17	14.5	0.12
IV	1 with 2	5.5	0.09	4.6	0.12	1.4	0.16	10.3	0.08
	1 with 3					1.2	0.14	12.8	0.10
	2 with 3					1.2	0.13	10.9	0.09
V	1 with 2	4.7	0.08	4.4	0.09	1.2	0.10	12.0	0.08
	1 with 3					0.9	0.07	13.7	0.08
	2 with 3					1.1	0.09	10.7	0.07
III to V	1 with 2			3.2	0.10	1.0	0.11	13.1	0.10
	1 with 3					1.0	0.10	13.7	0.10
	2 with 3					0.9	0.09	11.5	0.08
VI	1 with 2	5.5	0.07	6.3	0.10	1.3	0.10	9.1	0.05
	1 with 3					1.1	0.09	8.2	0.05
	2 with 3					1.1	1.08	9.5	0.05
VII	1 with 2	6.4	0.07	5.3	0.09	1.2	0.09	13.6	0.07
	1 with 3					1.4	0.10	17.0	0.09
	2 with 3					1.1	0.07	14.9	0.07
VIII	1 with 2	7.7	0.08	5.4	0.08	0.8	0.05	7.5	0.04
	1 with 3					0.9	0.06	7.5	0.04
	2 with 3					1.1	0.07	9.8	0.05
VI to VIII	1 with 2			6.2	0.10	1.1	0.08	12.0	0.07
	1 with 3					1.1	0.08	13.7	0.07
	2 with 3					1.1	0.08	11.4	0.06
III to VIII	1 with 2	5.3	0.07						

[a]Because of lack of space, the probable error of measurement has been abbreviated $P.\ E._M$ instead of $P.\ E._{Meas.}$

to VIII, the probable errors of measurement of the three tests which make up the Illinois Examination, also their ratios to the averages. Taking the entries in the first line of the table as an illustration, half of the differences between the intelligence scores actually obtained in Grade III with forms 1 and 2 and the theoretically true scores were found to be less than 3.5 points and half of them greater than this amount. For the arithmetic test half of these differences were less than 2.6 points; for the comprehension scores of the silent reading test half of the differences were less than 1.2, and for the rate scores half of the differences less than 13.7 words per minute. The probable errors of estimate, which are not given in Table VI, would of course be larger, that corresponding to the probable error of measurement of 13.7 just mentioned being about 18.5 words per minute.

It will be noticed that in Tables V and VI the columns containing the probable errors of estimate and of measurement, respectively, are followed by columns showing the ratios of these measures to the corresponding averages.[29] This is done because the mere statement of the size of a probable error of estimate or of measurement usually conveys little definite meaning unless one knows the size of the individual measures themselves. Just as an error of an inch is of slight significance in measuring the distance between two cities or even the length of a lot but is relatively significant in measuring a person's height and very important in fitting a piston to its cylinder, so an error of a given number of points on a test becomes more significant the smaller the score. It will be seen that whereas Tables V and VI show either a slight tendency for the probable errors to be greater in the higher grades or else no regular tendency at all, they reveal that relative to the average scores which increase from grade to grade, the errors become smaller, the ratios being considerably less in the eighth grade than in the third.

[29]There are certain objections raised to the use of these ratios which will not be discussed here, further than to admit that sometimes their use may be misleading. The writer believes, however, that in general their use is desirable both because the probable errors alone frequently convey little helpful information and because no better relative measure has been suggested.

CHAPTER III

THE COEFFICIENT OF CORRELATION

Definition of correlation. Before proceeding to discuss the use and interpretation of the coefficient of correlation it seems in order to define, first, what is meant by correlation in general, and, second, what is meant by the coefficient of correlation. Two characteristics or traits are said to be correlated when there is a tendency for changes in the value of one to be associated or occur concurrently with changes in the value of the other. If most of the changes in one of the things being dealt with are in the same direction as the corresponding changes in the other, the correlation is said to be positive or direct; if in opposite directions, it is said to be negative or inverse. For example, if pupils' marks in algebra and English are being correlated, and in most cases pupils who are relatively high in one are also relatively high in the other and likewise those who are low in one are generally low in the other, the correlation is positive; whereas, if pupils who stand high in algebra tend to rank low in English, and vice versa, it is negative. The greater the proportion of associated changes which are in the same direction, the greater is the amount of positive correlation; the greater the proportion in opposite directions, the greater the negative correlation. It is also true that the greater the agreement in relative magnitude of the concurrent changes, the greater the degree of correlation, whether positive or negative. For example, if a pupil who is 10 percent above the average in English is also 10 percent above the average in algebra, if one who is 5 percent above in English is 5 percent above in algebra, and so on for most of the cases, the correlation is higher than if this condition does not obtain.

Examples of both positive and negative correlation are very numerous and easily found. For example, it is usually found that the greater a person's height, the greater his weight; and that the older a child, the greater his strength. Therefore, height and weight, and children's age and strength are positively correlated. On the other hand, after an adult passes a certain age strength tends to decrease with advancing years so that the correlation is negative. This is also true when the two things compared are size of class and cost of instruction per pupil, since, on the whole, the larger the class the smaller is the cost for each member thereof.

The fact should be emphasized that the existence of correlation does not prove that there is any dependence or causal relationship between the two things correlated. It may be that such dependence exists, but it may also be that neither trait in any sense causes the other. Instead, the existing correlation may be due to the action of one or more outside factors which affect both the characteristics being dealt with. Sometimes the causal factor or factors may be even more remote than this, that is, some common cause may affect two characteristics or factors, each of these two may affect another, and so on, with the result that the final pair of characteristics considered, though relatively remote from the common cause, show correlation with each other. On the other hand, if the correlation between two traits is fairly high, the likelihood that one of them affects the other or that both are affected by a relatively proximate common cause, is great enough to be investigated as a probable hypothesis.

Definition of the coefficient of correlation. Although "coefficient of correlation" is sometimes used in a broad sense to include any one or all of a number of numerical expressions which summarize the degree of relationship between two variables, it is best to reserve this term for the product-moment coefficient of correlation, sometimes called the Pearson coefficient because its present extensive use is chiefly due to the English statistician, Karl Pearson. This expression, which is abbreviated by "r", is given by the formula:

$$r = \frac{\Sigma xy}{N\sigma_x \sigma_y},$$

in which x and y represent the deviations of the individual measures from their respective averages, and σ_x and σ_y the standard deviations of the distributions of the two variables. Σ is the sign of summation and N stands for the number of individuals. Therefore, what the formula accomplishes is to multiply the deviations or differences of each case from the averages, find the sum of these products for all the cases concerned, divide by the number of cases to find the average product, and further divide by the product of the two standard deviations in order to reduce the two distributions to a unit which is common and such that the value of the result cannot be greater than $\pm$ 1.00.

The coefficient of correlation is a measure of rectilinear or straight-line relationship only. That is, it measures the degree to which the data when tabulated in a correlation table approach a straight line or, in other words, the degree to which their graphical representation upon

TABLE VII. AN EXAMPLE OF A HIGH COEFFICIENT OF CORRELATION: THE CORRELATION BETWEEN POINT SCORES ON THE OTIS SELF-ADMINISTERING TEST, HIGHER EXAMINATION, FORM A AND INTELLIGENCE QUOTIENTS FOR A GROUP OF HIGH SCHOOL SENIORS

Point Scores	Intelligence Quotients.															T
	61–	66–	71–	76–	81–	86–	91–	96–	101–	106–	111–	116–	121–	126–	131–	
71–														1	9	10
66–												1	12	39	2	54
61–											1	30	89	9		129
56–										1	47	167	15	2		232
51–										74	264	14				352
46–									91	298	11					400
41–								87	245	8						340
36–							59	180	2							241
31–						1	23	79	4							107
26–						11	38	1	1							51
21–					5	14										19
16–				1	2											3
11–				2												2
6–																
1–		1														1
T	1		3	7	26	61	139	272	338	381	323	212	116	51	11	1941

$r = .97$

the X- and Y- axes[1] approaches a straight line. Tables VII and VIII, taken from an unpublished study made by the writer, are inserted to illustrate, respectively, rather close and very little approach to straight-line correlation. In Table VII the intelligence quotients of a number of high-school seniors are shown upon the X- or horizontal axis, and the point scores of the same seniors upon the Y- or vertical axis. This table shows, taking the top row as an example, that one senior who had a point score between 71 and 75, inclusive, had an intelligence quotient between 126 and 130, inclusive, and that nine seniors whose point scores were from 71 to 75 had intelligence quotients from 131 to 135. It can be seen that if a straight line were drawn diagonally through the table from the lower left-hand to the upper right-hand corner, the departure of the entries from it would be comparatively small. In other words the coefficient of correlation of .97 indicates rather close approach to perfect straight-line relationship. In Table VIII the horizontal or X-axis represents the number of semesters of Latin carried in high school,

[1]For further explanation of the X- and Y- axes see: Odell, op. cit., p. 37-38.

TABLE VIII. AN EXAMPLE OF A RATHER LOW COEFFICIENT OF CORRELATION: THE CORRELATION BETWEEN FRESHMAN LATIN MARKS AND NUMBER OF SEMESTERS OF LATIN CARRIED IN HIGH SCHOOL FOR A GROUP OF COLLEGE FRESHMEN

Freshman Latin Mark	Semesters of Latin in High School									T
	0	1	2	3	4	5	6	7	8	
96–	1				3	1		1	4	10
91–	1				3		4	1	8	17
86–	2		1		12		3	2	17	37
81–	5		1		8		1	1	12	28
76–	2		3		6	1	2		4	18
71–			1		2	1	1	1	3	9
66–	2				2					4
61–					1		$r = .31$			1
56–	3			1	1				1	6
T	16		6	1	38	3	11	6	49	130

and the vertical or Y- axis the freshman Latin mark in college. It shows, for example, that of the students whose freshman Latin marks were from 96 to 100, inclusive, one had carried no Latin in high school, three had carried four semesters, one five semesters, one seven semesters, and four ten semesters. An inspection of the table shows that, as is indicated by the coefficient of only .31, it is impossible to draw a straight line through it in such a direction that the entries show any considerable tendency to lie near this line.

It should also be noted that it is possible for two variables to be closely associated or correlated and yet show considerable departure from straight-line relationship. For example, if strength and age are compared throughout life, it is found that when persons are very young their strength is small, that as they become older, up to a certain limit, it increases, after which it decreases again. That is, there exists a fairly close relationship but not a rectilinear one. To completely measure such situations as this, an expression for curved-line or curvilinear relationship is needed. The one most commonly used is the ratio of correlation,[2] the discussion of which is outside the scope of this bulletin.

One may, for at least two reasons, think of the coefficient of correlation as a minimum measure of the amount of relationship existing. In

[2]See:
Odell, op. cit., p. 207-13.

the first place, if the association is rectilinear it measures all of it, but not more, whereas if the association is at all curvilinear it measures somewhat less than all of it. Secondly, in practically all cases variable or chance errors enter into the measurements of the two characteristics being correlated and the total effect of these errors, called attenuation,[3] is to make the computed or apparent coefficient of correlation less than the true one. If two series of similar measurements of each characteristic are available, it is possible to correct for the effect of such errors. When these are not available, all we can say is that the true value of r is as great as, or greater than, the one actually obtained.

As was stated above the value of the coefficient of correlation varies from + 1.00 through zero to — 1.00. A value of + 1.00 indicates perfect positive correlation, that is, that each score in one distribution deviates from its average in the same direction and by the same proportional amount as does the corresponding score in the other distribution. A coefficient of zero indicates that there is no correlation or, in other words, that the association between the two characteristics is purely a chance one. If r equals — 1.00, the two variables have perfect negative correlation, that is, each score in one series deviates from its average by the same proportional amount as the corresponding score of the other series, but in the opposite direction.

The coefficient of correlation as an index of the existence or absence of relationship. Two chief purposes for determining the value of the coefficient of correlation may be distinguished, although in many cases some element of both is present. One of these purposes is to learn whether there is any relationship at all between the two sets of paired facts under consideration. In other words, the question to which one is seeking an answer is "Are these two characteristics related to each other?" rather than, "How close is the relationship between these characteristics?" One may, for example, desire to determine whether or not any relationship exists between mental and physical ability, or between ability in arithmetical computation and in solving reasoning problems. In such cases, the fact that the coefficient of correlation is found to be appreciably greater than zero may be considered as evidence that there is some relationship present.

In interpreting a coefficient of correlation used for this purpose, one must know whether the obtained value of r is being considered as a measure of the correlation existing between the particular sets of paired facts for which it was computed, and these only, or whether these cases

[3] See:
Odell, op. cit., p. 181-85.

TABLE IX. CORRELATION BETWEEN SPEED AND QUALITY OF WRITING OF CHILDREN OF SCHOOL C

Grade	IV	V	VI	VII	VIII
Coefficient of Correlation	+.08(±.02)	−.10(±.04)	−14.(±.04)	−.34[a]	−.15(±.05)

[a]Freeman does not give the *P.E.* of this coefficient.

are to be considered merely as a sample of a larger number. If the latter is the case the obtained value is subject to errors of sampling just as is any other derived measure similarly used. As was explained earlier[4] in this bulletin, the most usual means of interpreting an obtained value of *r* when it is subject to errors of sampling is to compare it with its probable error. If, on the other hand, one is concerned merely with the cases actually measured and assumes that the measurements are accurate and the computations reliable, there is no need for interpreting *r* by comparison with such a measure of the reliability of sampling.

As was stated above, in many if not most cases in which the value of *r* is determined, the person or persons doing so do not make clear, and perhaps usually do not have definitely in mind, which one of the two purposes is predominant. In other cases, however, it is evident that one or the other is the more important in the particular situation. An example in which the purpose already described appears to be predominant may be found in an article by Freeman,[5] in which he gives coefficients of correlation between speed and quality of handwriting in Grades IV to VIII. Freeman's purpose was undoubtedly to present evidence as to whether, in general, there exists any relationship between quality and speed in handwriting, and therefore he gave the probable errors of all except the largest coefficient. It will be seen that the four

[4]See p. 21 for a discussion of the probable error of sampling.

[5]Freeman, F. N. "Some practical studies of handwriting." Elementary School Teacher, 14:167-79, December, 1913. Other examples of the same use of *r* may be found in the following references:

Starch, Daniel. Educational Psychology. New York: The Macmillan Company, 1923. p. 246.

Abernethy, Ethel M. "Correlations in physical and mental growth," Journal of Educational Psychology, 16:456-66. October, 1925.

Chapin, F. Stuart. "Extra-curricular activities of college students: a study in college leadership," School and Society, 23:212-16, February 13, 1926.

Furfey, Paul H. "Some preliminary results on the nature of developmental age," School and Society, 23:183-84. February 6, 1926.

small coefficients vary from two and one-half to four times their probable errors. Since it is customary to consider a value of r as fairly reliable if it is three times as large as its *P.E.* and to consider it as almost certainly reliable if it is four or five times its *P.E.*, it appears that the four small coefficients are probably reliable, but that only the one for Grade VII can certainly be said to be so. In other words, the data from Grades IV, V, VI and VIII indicate that, when all pupils are considered, speed and quality of handwriting probably are slightly negatively associated, whereas those for Grade VII seem to show that there is no doubt of the existence of negative relationship.

To illustrate further the interpretation of r when used for this purpose, let us suppose that the correlation between the average number of pupils to the teacher and the average salaries of teachers has been found for all cities of more than 100,000 population and also for a random sample of fifty cities having from 50,000 to 100,000 population. Furthermore, let us suppose that in both cases the obtained value of r is .20. For the cities above 100,000 such a value would indicate definitely that there was a real, though small, relationship between the average number of pupils and average salaries. This is true because all cities of the class were included and there was no sampling. On the other hand, for the cities of from 50,000 to 100,000 the value of r is subject to a *P.E.* of .09[6] because in this case a sampling was made. Therefore, it is probable but by no means certain that some correlation really exists for all cities of this size, the chances being about 6 to 1 in its favor. If a sample of one hundred instead of fifty cities were taken, the *P.E.* of r would be reduced from .09 to .06 and we might say it was fairly certain that the correlation for the whole group was positive, since the chances are about 40 to 1 in its favor.[7] If the size of the sample was increased still more the probable error of r would, of course, be still further reduced and the chances that r is certainly reliable proportionately increased.

The coefficient of correlation as a measure of the closeness of relationship or reliability of prediction. The second purpose for which the coefficient of correlation is commonly used is to indicate just how close is the relationship or association between two characteristics or how accurately one can be estimated or predicted when the other is known. This purpose really assumes that there is some definite rela-

[6]The formula for the probable error of the coefficient of correlation is $.6745\frac{1-r^2}{\sqrt{N}}$. Hence, in this case $P.E._r = .6745\frac{1-.20^2}{\sqrt{50}} = .09$.

[7]See p. 14 of this bulletin.

tionship, either positive or negative, and seeks to determine how nearly it approaches complete or perfect association or, in other words, what the probable accuracy of estimating one variable from the other is. The value of *r* gives a measure of the accuracy or reliability of the prediction or prognosis possible. For example, if we know that the coefficient of correlation between height and weight is .75, we have some idea as to how closely a given person's height can be estimated if his weight is known, or vice versa. How definite an idea we have depends largely upon the amount of our experience in meeting and dealing with similar situations involving various values of the coefficient. The name "coefficient of reliability" or "coefficient of self-correlation" is frequently applied to the coefficient of correlation between two series of duplicate measurements of the same individuals, such as those yielded by duplicate forms of a test or measurements of height by two persons. Sometimes these names are also applied to coefficients of correlation between two series of similar but not duplicate measurements, such as those yielded by two different intelligence tests or by two reading tests. For example, if pupils' abilities in the fundamental operations of arithmetic are measured by Form 1 of the Courtis Research Tests in Arithmetic, Series B, and later this is repeated, or one of the equivalent forms used, the coefficient of correlation between the two series of scores is the coefficient of reliability. Likewise, the term is often though less frequently applied to the correlation between the scores of a group of pupils on, for example, the National Intelligence Tests and the Illinois General Intelligence Scale.

An example of the use of the coefficient of correlation with this purpose predominating is shown by Table X, prepared by Starch.[8] It contains the coefficients of correlation found between school marks of two groups of pupils in various subjects. This table shows, for example, that a pupil's grade in arithmetic can be more closely predicted from

---

[8] Starch, Daniel. "Correlation among abilities in school studies, Journal of Educational Phychology, 3:415-18, September, 1913. Other examples may be found by consulting the following:

Orleans, J. S. "The ability to spell," School and Society, 23:407-08, March 27, 1926.

Nanninga, S. P. "A critical study of rating traits," Educational Administration and Supervision, 12:114-19, February, 1926.

Hull, C. L., and Limp, C. E. "The differentiation of the aptitudes of an individual by means of test batteries," Journal of Educational Psychology, 16:73-88, February, 1925.

Ruch, G. M., and Stoddard, G. D. "Comparative reliabilities of five types of objective examinations," Journal of Educational Psychology, 16:89-103, February, 1925.

TABLE X. COEFFICIENTS OF CORRELATION BETWEEN MARKS OF TWO GROUPS OF PUPILS IN SEVERAL SCHOOL SUBJECTS

	First Group	Second Group
Arithmetic and language	.73	.85
Arithmetic and geography	.74	.83
Arithmetic and history	.73	...
Arithmetic and reading	.45	.67
Arithmetic and spelling	.42	.55
Language and geography	.86	.85
Language and history	.77	...
Language and reading	.80	.83
Language and spelling	.77	.71
Geography and history	.81	...
Geography and reading	.83	.80
Geography and spelling	.68	.52
History and reading	.67	...
History and spelling	.37	...
Reading and spelling	.72	.58

his grade in language, with which the correlations are .73 and .85, than from his grade in spelling, which correlates with it only .42 and .55.

In connection with the use of r for this purpose one should bear in mind that its value may be large enough to indicate that there is a definite association between the two characteristics correlated and yet not large enough to enable one to place much confidence in the prediction of the probable amount of one trait possessed by an individual when that of the other is known. Furthermore, the value of r in itself does not give a direct measure of the size of the errors liable to be present in predictions or estimates based upon the data from which r was computed. It is, therefore, frequently desirable to interpret coefficients of reliability and other coefficients of correlation used for estimating one characteristic from another by finding the probable errors of estimate and of measurement[8a] associated with them. The following paragraph will describe the method of doing so.

Interpretation of the coefficient of correlation in terms of the probable errors of estimate and of measurement. The formulae for the probable errors of estimate and of measurement which were given on p. 29 show that their magnitude depends upon two things—the coefficient of correlation and the median deviation of the distribution. We can therefore easily find for any given value of r the size of the probable errors of estimate and of measurement in terms of $Md.D.$ as the

[8a] For a discussion of these measures see above, p. 28 et seq.

TABLE XI. VALUES OF THE PROBABLE ERRORS OF ESTIMATE AND OF MEASUREMENT CORRESPONDING TO CERTAIN VALUES OF THE COEFFICIENT OF CORRELATION

Coefficient of Correlation	Probable Error of Estimate	Probable Error of Measurement
1.00	.0000 *Md. D.*	.0000 *Md. D.*
.99	.1411 *Md. D.*	.1000 *Md. D.*
.98	.1990 *Md. D.*	.1414 *Md. D.*
.97	.2431 *Md. D.*	.1732 *Md. D.*
.96	.2800 *Md. D.*	.2000 *Md. D.*
.95	.3122 *Md. D.*	.2236 *Md. D.*
.90	.4359 *Md. D.*	.3162 *Md. D.*
.80	.6000 *Md. D.*	.4472 *Md. D.*
.70	.7141 *Md. D.*	.5477 *Md. D.*
.60	.8000 *Md. D.*	.6325 *Md. D.*
.50	.8660 *Md. D.*	.7071 *Md. D.*
.40	.9165 *Md. D.*	.7746 *Md. D.*
.30	.9539 *Md. D.*	.8367 *Md. D.*
.20	.9798 *Md. D.*	.8944 *Md. D.*
.10	.9950 *Md. D.*	.9487 *Md. D.*
.00	1.0000 *Md. D.*	1.0000 *Md. D.*

unit. Table XI has been inserted to give the probable errors of estimate and of measurement for the values of r from .00 to .90 at intervals of .10, and from .95 to 1.00 at intervals of .01. For example, if the coefficient of correlation is .99 the probable error of estimate is .1411 *Md.D.* and that of measurement .1000 *Md.D.* Similarly, if $r = .70$, $P.E._{Est.} = .7141$ *Md.D.* and $P.E._{Meas.} = .5477$ *Md.D.* Glancing over the whole table one sees that an increase of the same amount in the coefficient of correlation produces a greater decrease in the errors when r is high than when it is low.

The preceding discussion has probably not made absolutely clear the significance of a probable error of estimate or of measurement expressed in terms of *Md.D.* The following statement may be helpful in this connection. When $r = .00$, or in other words when no correlation at all exists, both $P.E._{Est.}$ and $P.E._{Meas.} = 1.0000$ *Md.D.* This means that if one attempted to estimate scores in one distribution from those in the other by making pure guesses,[9] he might expect that in half of his esti-

[9] In using the term "pure guesses" it is understood that the person so guessing knows the limits and the general shape of the distribution of scores being guessed. He does not, however, have at his command any information whatsoever which helps him in guessing the location of any particular score within this distribution.

mates or guesses he would be in error by amounts less than *Md.D.*, and in half by amounts greater. From this point of view it can be seen that even though the coefficient of correlation is rather large a great deal of the guessing element is present in estimating scores in one distribution from those in the other. Many people commonly think that if $r = .85$ or .90 the association is very close or almost perfect, whereas as a matter of fact an estimate is still half a pure guess when $r = .866$, and even when $r = .968$ an estimate is one-fourth a pure guess. When one is estimating a true score from a score actually obtained the estimate is half a guess when $r = .75$, and one-fourth a guess when r is almost .94. Thus it can be seen that the coefficient of correlation must either approach 1.00 very closely, or equal it, before the errors of estimate and of measurement are small enough to be negligible. On the other hand, even when these errors are considerable they are less than would be the case if no correlation existed, and therefore one can make better estimates of scores in one distribution from those in another if there is any correlation at all between the two than he can if no helpful information of any kind is available.[10]

To illustrate still more clearly the meaning of the probable errors of estimate and of measurement Figure 5[11] is given. It shows the correlation between scores on Forms 1 and 2 of the Illinois General Intelligence Scale. This figure is in general similar to Tables VII and VIII except that instead of numbers to show how many scores fall in each cell it contains a dot for each score. The height of each dot above the base line (X-axis) shows the Form 1 score made by the individual represented by the dot, and its distance to the right of the vertical line at the left of the table (Y-axis) shows the Form 2 score of the same individual. In a few cases figures showing these distances have been placed in parenthesis after the dots. In such cases the first of these two numbers indicates the Form 2 score or X distance and the latter of the two the Form 1 score or Y distance. For example, near the upper right hand corner of the figure is a dot representing a pupil who made a score of 104 on Form 2 and 118 on Form 1.

---

[10]The actual procedure of estimating scores in one series from those in another with which it is correlated involves the use of the regression equation, which is based upon the averages and standard deviations of the two series and the coefficient of correlation. For an explanation of regression, see:

ODELL, op. cit., p. 189-96.

[11]This figure is taken from:

MONROE, W. S. "The Illinois examination." University of Illinois Bulletin, Vol. 19, No. 9. Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921. 45 p.

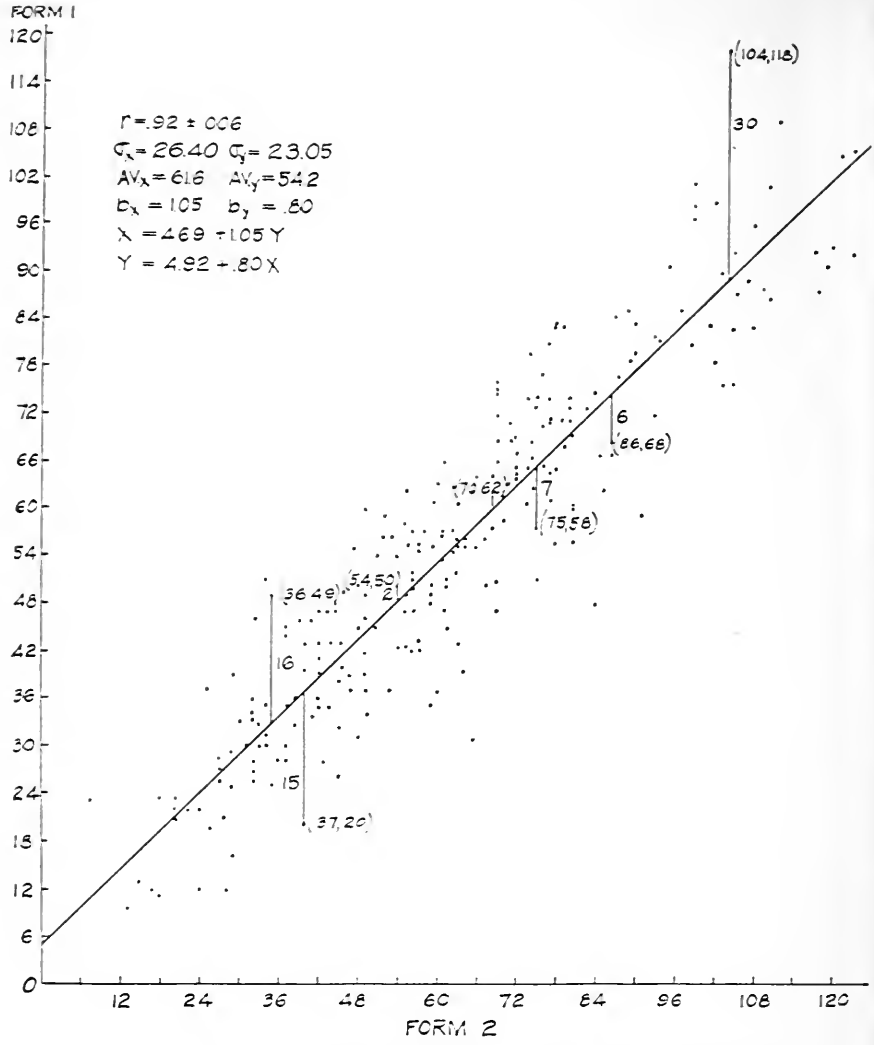

FIGURE 5. CORRELATION OF FORM 1 SCORES WITH FORM 2 SCORES OF THE ILLINOIS GENERAL INTELLIGENCE SCALE, FIFTH GRADE

The long diagonal line in the figure is the Y regression line. In other words it is a graphical representation of the best possible straight-line equation for estimating the Y or Form 1 score when the X or Form 2 score is known, which can be derived from the correlation between the two distributions and their spread or variability. If the correlation were perfect there would be no errors in such estimates and all the dots

would lie upon the diagonal line. As it is the vertical distance from each dot to this line represents the error involved in estimating the Form 1 score from the Form 2 score. In several cases the vertical lines connecting the dots with the diagonal have been drawn in, with a number beside each line to indicate its length, or in other words the error of estimate in that particular case. For example, the error of estimate for the case previously mentioned as having scores of 104 and 118 is 30. Substituting in the formula given above,[12] it is found that $P.E._{Est.}$ is 6.06. Therefore we know that the errors involved in estimating Form 1 from Form 2 scores are less than this in half of the cases and greater in the other half. In other words, the vertical distances from half of the dots to the diagonal line are less than 6.06, whereas those from the other half are greater than this amount. A similar diagonal line representing the estimates of Form 2 scores from Form 1 scores could also be drawn. If it were, the horizontal distances from the points to that line would represent the errors in those estimates.

Interpretation of the coefficient of correlation by comparison with the sizes of coefficients of correlation commonly found. A second means of interpreting r is by comparing its value with the coefficients of correlation found to exist in certain relatively familiar situations. One has a more or less definite idea of the extent to which tall people tend to weigh more and short people less, of how much children tend to resemble their parents in height, and so on. Therefore, by comparing the value of a coefficient of correlation with those usually found in some of these fairly common and well-known cases, such an idea may be formed as, for example, that the relationship in question is somewhat closer than that between height and weight, or about the same as that between school marks in Latin and in French. It is true that our ideas as to just how closely two characteristics, even though they are very common, are related, are decidedly subjective and therefore often considerably in error, yet such comparisons have some value in interpreting coefficients of correlation. To aid the reader in making such interpretations the following table showing the magnitude of the coefficients of correlation usually found between pairs of certain fairly common characteristics is given.

Interpretation of r in terms of displacement. Another means of interpreting the value of r is in terms of the differences in individuals' relative positions in the two series of measures correlated or, as this is commonly called, in terms of displacement. For example, if height and

[12] See p. 29.

TABLE XII. SIZES OF COEFFICIENTS OF CORRELATION COMMONLY FOUND

Cost of instruction with total cost of education	.90–.95
First and second applications of an individual intelligence test	.90–.95
Ages of husbands with ages of wives	.85–.95
First and second applications of a standardized group test	.60–.90
School marks in subjects supposed to be more or less akin, such as English and foreign language, or mathematics and physics	.40–.70
Heights of fathers with heights of sons	.40–.60
School marks in subjects supposed to have little in common, such as Latin and domestic science, or English and manual training	.30–.50
Quality of handwriting and intelligence test scores	.00–.10

weight are the two characteristics concerned and a certain individual is fourth from the top in height, his displacement is the amount that he differs from this position in weight. The interpretation is usually made by expressing the probability that position or rank in one variable does not differ by more than a certain distance or number of places from that in the other. A table which may be used for this purpose has been prepared and published by Otis,[13] but is not reproduced here because the interpretation of *r* through the amount of displacement has not come into common use. Also it is somewhat difficult to comprehend readily just what is meant by this method of interpretation.

The interpretation of the coefficient of correlation in terms of adjectives. Some writers have undertaken to define the meaning of coefficients of correlation by means of certain adjectives which they apply to coefficients of various sizes. Rugg,[14] for example, states that his experience "has led him to regard correlation as 'negligible' or 'indifferent' when *r* is less than .15 to .20; as being 'present but low' when *r* ranges from .15 or .20 to .35 or .40; as being 'markedly present' or 'marked' when *r* ranges from .35 or .40 to .50 or .60; as being 'high' when it is above .60 or .70. With the present limitations on educational testing few correlations in testing will run above .70, and it is safe to regard this as a very high coefficient." McCall[15] likewise offers a statement of this sort, but briefer than Rugg's, as follows:

---

[13]Otis, A. S. Statistical Method in Educational Measurement. Yonkers: World Book Company, 1925, p. 225.

[14]Rugg, H. O. Statistical Methods Applied to Education. Boston: Houghton Mifflin Company, 1917, p. 256. Also see:

Rugg, H. O. A Primer of Graphics and Statistics for Teachers. Boston: Houghton Mifflin Company, 1925, p. 97.

[15]McCall, W. A. How to Measure in Education. New York: The Macmillan Company, 1922, p. 392-93.

"when r is 0 to $\pm$ .4 correlation is low, or
$\pm$ .4 to $\pm$ .7 correlation is substantial, or
$\pm$ .7 to $\pm$ 1.0 correlation is high."

The chief purpose of the present writer in mentioning this method of interpreting r is to point out that the use of adjectives is decidedly unsatisfactory and indeed often meaningless unless they are employed in a definitely limited situation. Whether a coefficient of correlation is high or fair or low depends upon the purpose for which it is employed and the data for which it is computed. A coefficient of .30 or .40, for example, is high enough to indicate that there is definite relationship between the two things correlated but at the same time it is so low that, as has been shown, estimates of one of the traits from the other are scarcely better than mere guesses. Again, a correlation of .80 between school marks in chemistry and in physics is relatively high since the usual correlation between such marks is considerably lower than this, but a correlation of the same size between two applications of the same individual intelligence test is low since the best of such tests yield correlations of .90, or above. The writer therefore wishes to repeat that it is very undesirable to describe the amount of correlation by means of adjectives unless it is done in view of a definite and particular situation.

Effect of spread of data upon value of r. Another topic that should be treated is the interpretation of the coefficient of correlation in view of certain facts concerning the data for which it is computed. One of these important facts which should be known is the spread of data, that is, the extent to which they vary or scatter away from their average. Probably the most common occasion on which this is important is when there is a difference in the number of school grades that contributed the data for the two or more correlations being compared. Frequently, coefficients of correlation are determined between series of measurements obtained from a single grade, whereas on other occasions they are based upon those from several grades. In many cases the spread of the characteristic measured increases as the number of grades is increased. For example, the variations in height, weight, mental age, score upon a subject-matter test, and so forth, may be expected to be greater in two grades than in one, greater in three than in two, and so on. The effect of this increased spread is to raise the obtained value of the coefficient of correlation although, of course, the degree of relationship is not changed. For example, the correlation between age and height may average only .40 in each grade but if all grades from one to eight are included, it will probably be at least .70 or .80. Sometimes the effect of increasing the spread is so pronounced that correlations which are nega-

tive for a single grade or other limited group become positive for a more scattered or variable group. One of the common examples of this is the correlation between chronological and mental age. Within any given grade it is almost always true that the younger pupils are the brighter and the older ones the duller, so that the correlation between mental and chronological age in a single grade is almost always negative, generally from — .20 to — .50. If two or three grades are taken together this correlation usually changes to about zero, whereas if five or six are included it becomes positive, probably from .50 to .70, agreeing with the fact that older children tend to have higher mental ages than younger ones.

Because of this effect of the spread of the group upon the value of *r*, any given value thereof should be accompanied by a statement defining the group for which it was computed. It is also frequently desirable to give some measure, such as the median or standard deviation, of the spread of each group. By means of a formula[16] not given here, one can then make allowance for the effect of different degrees of spread upon the coefficient of correlation, and thus compare different values of *r* upon a true basis.

It should be noted that one group may have a greater spread than another in some characteristic or characteristics other than those correlated without affecting the value of *r*. For example, the range or spread of intelligence quotients in an average group of pupils from several grades is little, if any, greater than in a group from one grade only, altho the spread of the pupils as regards grade, age, and so on, is much greater in the first group. The same is true of the school marks given by teachers, of health ratings of teachers' salaries, and so on. Therefore, in cases such as these it is not necessary to allow for the fact that several grade groups instead of one are included. For example, the correlation between *I.Q.*'s and school marks for a group from several grades would be practically the same as for a single grade group. On the other hand, it may be that there is a difference in the spread of these characteristics due to some less common basis of grouping than grades. If, for example, pupils have been grouped according to their mental ability the spread of *I.Q.*'s will be greater in a combined group embracing sections of various abilities than in a single group of bright, average, or dull pupils.

Averaging coefficients of correlation. One not infrequently sees such a statement as that the average coefficient of correlation is .60 or

[16]This is given in:
Odell, op. cit., p. 174-77.

.85, for example. The process of averaging coefficients of correlation is not one which should be indulged in without taking precautions that the average so obtained is statistically justified. To illustrate this, if the correlation between intelligence and reading ability is .40 for one group of pupils and .60 for another, it is only by chance that it will be .50 if the data for the two groups are thrown into one correlation table. To insure this result the numbers of cases in the two groups, the averages and the spreads of the two groups around their averages, must be the same. These conditions are very rarely fulfilled. For practical purposes, however, if the averages and spreads are not very different and if each correlation is weighted by the number of cases which contribute to it, the average obtained may be considered as fairly representative. It is, however, usually if not always much better to give all the obtained values of r than to give merely their average. Certainly, if the average is given it should be made clear that it is only a more or less rough or approximate estimate of the amount of correlation.